

PATENT ABSTRACTS OF JAPAN

(11)Publication number : 2002-108846

(43)Date of publication of application : 12.04.2002

(51)Int.Cl.

G06F 17/21

G06K 9/03

G06K 9/20

H04N 1/387

(21)Application number : 2000-294877

(71)Applicant : FUJI XEROX CO LTD

(22)Date of filing : 27.09.2000

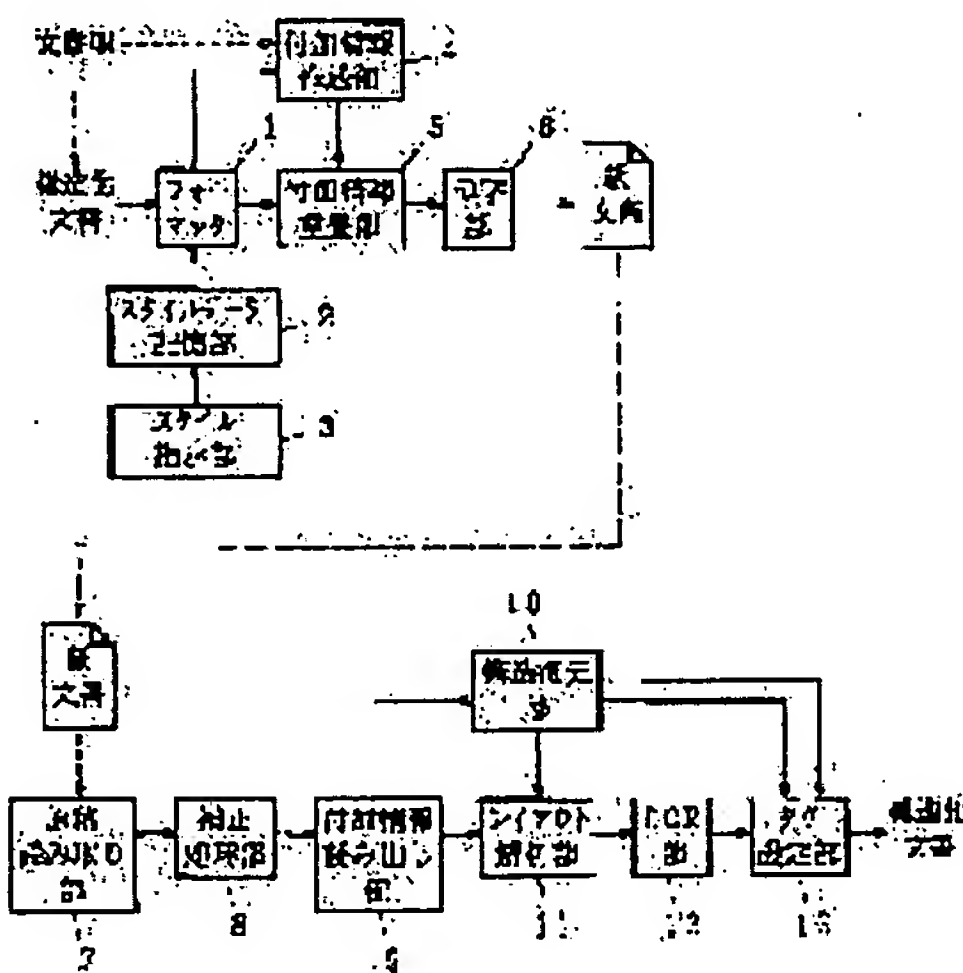
(72)Inventor : SUZUKI KAZUHIRO

(54) DEVICE/METHOD FOR PROCESSING DOCUMENT IMAGE AND RECORDING MEDIUM

(57)Abstract:

PROBLEM TO BE SOLVED: To provide a document image processor capable of restoring structured document after turning the structured document into a printable image.

SOLUTION: When a formatter 1 formats the structured data in accordance with style data, the formatter 1 sends its positional information to an additive information preparing part 4. The part 4 makes the positional information correspondent to document structure to turn the information and a document type to be additive information, an additive information superimposing part 5 superimposes it to a document image, and a printing part 6 prints it. A printed paper document is read by a document reading part 7. After it is correctly processed by a correcting processing part 8, the additive information is separated by an additive information reading part 9. A structure restoring part 10 restores document structure based on the additive information. A layout analytic part 11 analyzes layout to segment a document element and recognizes a text by an OCR part 12. Then, a tag setting part 13 inserts the document element to the document structure restored by the part 10 to restore the structured document.



LEGAL STATUS

[Date of request for examination]

[Date of sending the examiner's decision of rejection]

[Kind of final disposal of application other than the examiner's decision of rejection or application converted registration]

[Date of final disposal for application]

[Patent number]

[Date of registration]

[Number of appeal against examiner's decision of rejection]

[Date of requesting appeal against examiner's decision of rejection]

[Date of extinction of right]

(19) 日本国特許庁 (J P)

(12) 公開特許公報 (A)

(11) 特許出願公開番号
特開2002-108846

(P 2 0 0 2 - 1 0 8 8 4 6 A)

(43) 公開日 平成14年4月12日 (2002.4.12)

(51) Int. Cl. ⁷	識別記号	F I	テ-マコ-ト (参考)	
G06F 17/21	530	G06F 17/21	530	T 5B009
	566		566	P 5B029
G06K 9/03		G06K 9/03		J 5B064
9/20	340	9/20	340	C 5C076
			340	B

審査請求 未請求 請求項の数22 O L (全15頁) 最終頁に続く

(21) 出願番号 特願2000-294877 (P 2000-294877)

(22) 出願日 平成12年9月27日 (2000.9.27)

(71) 出願人 000005496

富士ゼロックス株式会社

東京都港区赤坂二丁目17番22号

(72) 発明者 鈴木 一弘

神奈川県海老名市本郷2274番地 富士ゼロックス株式会社内

(74) 代理人 100101948

弁理士 柳澤 正夫

F タ-ム (参考) 5B009 NA05 NA06

5B029 AA01 BB02 CC26 CC27

5B064 AA01 BA01 CA08 FA07 FA15

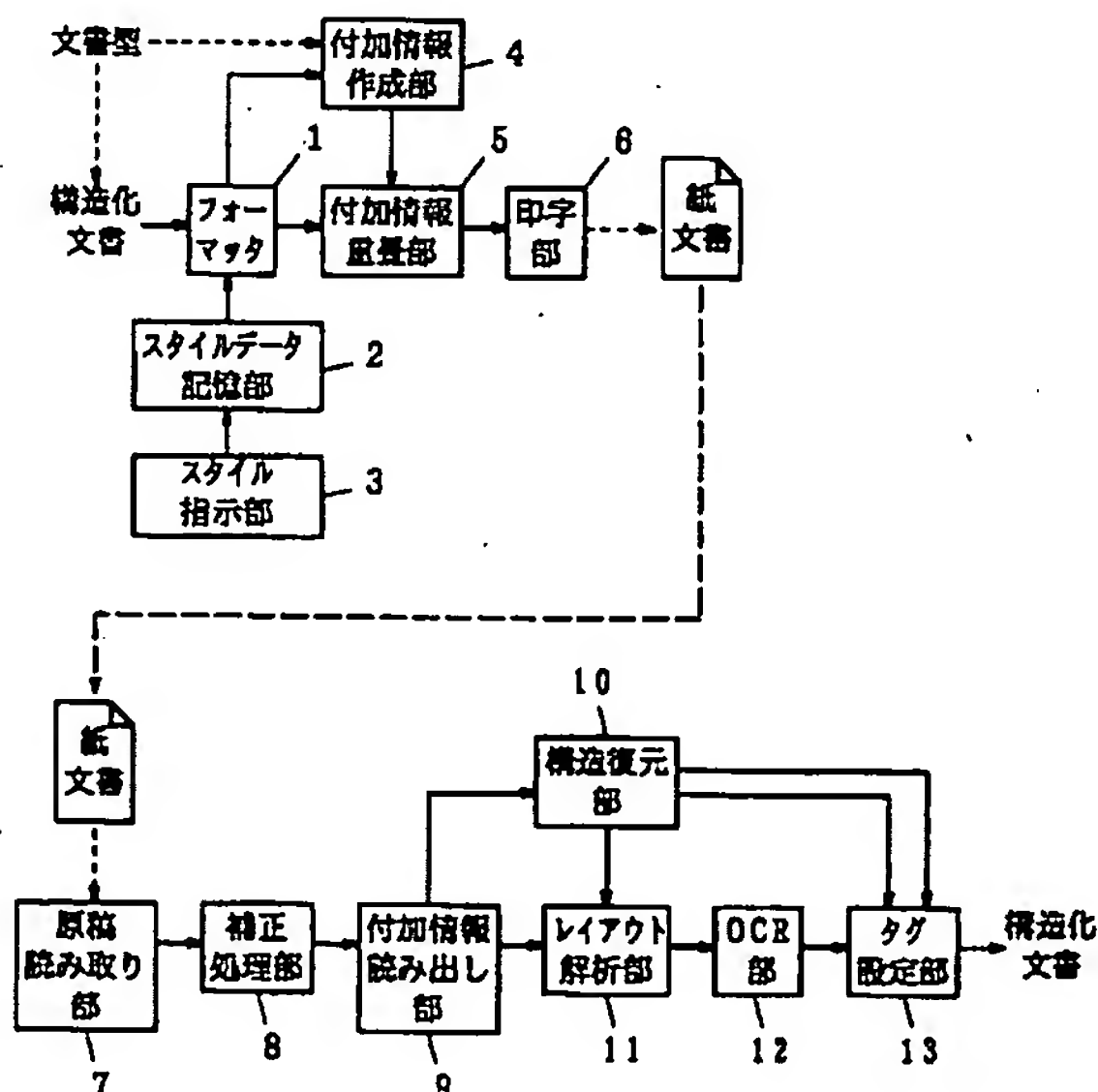
5C076 AA14 AA16 AA17 BA06

(54) 【発明の名称】 文書画像処理装置、文書画像処理方法、および記録媒体

(57) 【要約】

【課題】 構造化文書を印刷可能なイメージとした後に、構造化文書を復元することができる文書画像処理装置を提供する。

【解決手段】 フォーマッタ1がスタイルデータに従って構造化文書をフォーマットするときに、その位置情報を付加情報作成部4に伝える。付加情報作成部4は、その位置情報と文書構造とを対応付け、その情報及び文書型を付加情報として付加情報重畳部5で文書イメージに重畳して印字部6で印字する。印字された紙文書は原稿読み取り部7で読み取り、補正処理部8で補正処理後、付加情報読み出し部9で付加情報を分離する。構造復元部10は、付加情報をもとに文書構造を復元する。一方、レイアウト解析部11はレイアウト解析して文書要素を切り出し、テキストをOCR部12で認識する。そして、タグ設定部13で文書要素を構造復元部10で復元した文書構造に挿入し、構造化文書を復元する。



【特許請求の範囲】

1
【請求項 1】 論理構造を持った構造化文書にレイアウトを施して印刷可能とする文書画像処理装置において、論理構造に対応する体裁情報を定義したスタイルデータを記憶するスタイルデータ記憶手段と、前記スタイルデータに基づいて構造化文書にレイアウトを施して印刷処理に適合した第 1 の印刷イメージに成形するとともに前記構造化文書の文書要素の意味もしくは構造を識別するための識別子と前記文書要素が前記第 1 の印刷イメージ内に占める領域情報を生成する成形手段と、前記論理構造を表す論理構造情報に加えて前記構造化文書の文書要素の意味もしくは構造を識別するための識別子と前記文書要素が前記第 1 の印刷イメージ内に占める領域情報の対応関係を示す付加情報を生成する付加情報作成手段と、前記第 1 の印刷イメージに前記付加情報を重畳して第 2 の印刷イメージを生成する付加情報重畳手段を備えたことを特徴とする文書画像処理装置。

2
【請求項 2】 構造化文書の論理構造を表す論理構造情報に加えて前記構造化文書の文書要素の意味もしくは構造を識別するための識別子と前記文書要素が印刷イメージ内に占める領域情報の対応関係を示す付加情報が重畳された印刷イメージを読み取る文書画像処理装置において、前記印刷イメージを読み取り画像情報に変換する読み取り手段と、前記画像情報から前記付加情報を読み出す付加情報読み出し手段と、前記付加情報から前記論理構造情報を得て前記論理構造を復元するとともに前記構造化文書の文書要素の意味もしくは構造を識別するための識別子と前記文書要素が前記印刷イメージ内に占める領域情報の対応関係を得て前記識別子と前記識別子に対応する前記領域情報を出力する構造復元手段と、前記画像情報のレイアウトを認識して文字情報の含まれる領域である文字ブロックを抽出して前記領域情報に対応する前記文字ブロックを出力するレイアウト解析手段と、前記文字ブロック内の文字を認識してテキストデータを出力する文字認識手段と、前記論理構造を得るとともに前記識別子に対応する文書要素位置に前記テキストデータを挿入して構造化文書を得るタグ設定手段を備えたことを特徴とする文書画像処理装置。

3
【請求項 3】 さらに、前記読み取り手段で読み取った前記画像情報に対して傾き補正または天地補正を施して補正画像情報を得る補正手段を有し、前記付加情報読み出し手段は、前記補正画像情報から前記付加情報を読み出すものであり、前記レイアウト解析手段は、前記補正画像情報のレイアウトを認識するものであることを特徴とする請求項 2 に記載の文書画像処理装置。

4
【請求項 4】 前記論理構造情報は、構造化文書の文書型であることを特徴とする請求項 1 ないし請求項 3 のいずれか 1 項に記載の文書画像処理装置。

5
【請求項 5】 前記論理構造情報は、構造化文書の特定の文書型に対応付けられた識別情報であることを特徴と

する請求項 1 ないし請求項 3 のいずれか 1 項に記載の文書画像処理装置。

6
【請求項 6】 前記印刷イメージ内に占める前記領域情報は、前記印刷イメージ内に設定された座標系によって記述されることを特徴とする請求項 1 ないし請求項 3 のいずれか 1 項に記載の文書画像処理装置。

7
【請求項 7】 前記印刷イメージ内に占める前記領域情報は、前記印刷イメージ内の領域に対して一意に識別可能な符号を所定の手順で割り当てたものであることを特徴とする請求項 1 ないし請求項 3 のいずれか 1 項に記載の文書画像処理装置。

8
【請求項 8】 前記構造化文書の文書要素の意味もしくは構造を識別するための識別子と前記文書要素が前記第 1 の印刷イメージ内または前記印刷イメージ内に占める領域情報の対応関係は、印刷イメージの生成側と読取側とで共有していることを特徴とする請求項 1 ないし請求項 3 のいずれか 1 項に記載の文書画像処理装置。

9
【請求項 9】 前記構造化文書の文書要素の意味もしくは構造を識別するための識別子と前記文書要素が前記印刷イメージ内に占める領域情報の対応関係は、前記論理構造を階層表記し、前記文書要素を前記第 1 の印刷イメージ内に占める領域情報に置き換えて記述したものであることを特徴とする請求項 1 ないし請求項 3 のいずれか 1 項または請求項 6 または請求項 7 に記載の文書画像処理装置。

10
【請求項 10】 論理構造を持った構造化文書にレイアウトを施して印刷可能とする文書画像処理装置において、前記論理構造に対応してあらかじめ定義されたスタイルデータを記憶したスタイルデータ記憶手段と、前記スタイルデータに基づいて構造化文書にレイアウトを施して第 1 の印刷イメージに成形するとともに前記論理構造と前記レイアウト情報の組み合わせを同定する組み合わせ情報を生成する成形手段と、前記組み合わせ情報を付加情報とする付加情報作成手段と、前記第 1 の印刷イメージに前記付加情報を重畳して第 2 の印刷イメージを生成する付加情報重畳手段を備えたことを特徴とする文書画像処理装置。

11
【請求項 11】 構造化文書の論理構造とレイアウト情報の組み合わせを同定する組み合わせ情報が付加情報として重畳された印刷イメージを読み取る文書画像処理装置において、前記印刷イメージを読み取り画像情報に変換する読み取り手段と、前記画像情報から前記付加情報を読み出す付加情報読み出し手段と、前記付加情報から前記組み合わせ情報を得て予め設定されている前記論理構造と前記レイアウト情報の定義に基づいて前記構造化文書の文書要素の意味もしくは構造を識別するための識別子と前記文書要素が前記印刷イメージ内に占める領域情報の対応関係を得て前記識別子と前記識別子に対応する前記領域情報を出力する構造復元手段と、前記画像情報のレイアウトを認識して文字情報の含まれる領域であ

る文字ブロックを抽出し前記領域情報に対応する前記文字ブロックを出力するレイアウト解析手段と、前記文字ブロック内の文字を認識してテキストデータを出力する文字認識手段と、前記論理構造を得るとともに前記識別子に対応する文書要素位置に前記テキストデータを挿入して前記構造化文書を得るタグ設定手段を備えたことを特徴とする文書画像処理装置。

【請求項 1 2】 さらに、前記読み取り手段で読み取った前記画像情報に対して傾き補正または天地補正を施して補正画像情報を得る補正手段を有し、前記付加情報読み出し手段は、前記補正画像情報から前記付加情報を読み出すものであり、前記レイアウト解析手段は、前記補正画像情報のレイアウトを認識するものであることを特徴とする請求項 1 1 に記載の文書画像処理装置。

【請求項 1 3】 前記付加情報は、前記印刷イメージに可視に重畳されていることを特徴とする請求項 1 ないし請求項 1 2 のいずれか 1 項に記載の文書画像処理装置。

【請求項 1 4】 前記付加情報は、前記印刷イメージに不可視に重畳されていることを特徴とする請求項 1 ないし請求項 1 2 のいずれか 1 項に記載の文書画像処理装置。

【請求項 1 5】 論理構造を持った構造化文書にレイアウトを施して印刷可能とする文書画像処理方法において、論理構造に対応する体裁情報を定義したスタイルデータに基づいて構造化文書にレイアウトを施して印刷処理に適合した第 1 の印刷イメージに成形するとともに、前記構造化文書の文書要素の意味もしくは構造を識別するための識別子と前記文書要素が前記第 1 の印刷イメージ内に占める領域情報を生成し、前記論理構造を表す論理構造情報に加えて前記構造化文書の文書要素の意味もしくは構造を識別するための識別子と前記文書要素が前記第 1 の印刷イメージ内に占める領域情報の対応関係を示す付加情報を生成して前記第 1 の印刷イメージに重畳し、第 2 の印刷イメージを生成することを特徴とする文書画像処理方法。

【請求項 1 6】 構造化文書の論理構造を表す論理構造情報に加えて前記構造化文書の文書要素の意味もしくは構造を識別するための識別子と前記文書要素が印刷イメージ内に占める領域情報の対応関係を示す付加情報が重畳された印刷イメージを読み取る文書画像処理方法において、前記印刷イメージを読み取って画像情報に変換し、前記画像情報から前記付加情報を読み出し、前記付加情報から前記論理構造情報を得て前記論理構造を復元するとともに、前記構造化文書の文書要素の意味もしくは構造を識別するための識別子と前記文書要素が前記印刷イメージ内に占める領域情報の対応関係を得て前記識別子と前記識別子に対応する前記領域情報を取得し、前記画像情報のレイアウトを認識して文字情報の含まれる領域である文字ブロックを抽出し、前記領域情報に対応する前記文字ブロック内の文字を認識してテキストデー

タとし、前記論理構造を得るとともに前記識別子に対応する文書要素位置に前記テキストデータを挿入して構造化文書を得ることを特徴とする文書画像処理方法。

【請求項 1 7】 論理構造を持った構造化文書にレイアウトを施して印刷可能とする文書画像処理方法において、前記論理構造に対応してあらかじめ定義されたスタイルデータに基づいて構造化文書にレイアウトを施して第 1 の印刷イメージに成形するとともに、前記論理構造と前記レイアウト情報の組み合わせを同定する組み合わせ情報を生成して付加情報とし、該付加情報を前記第 1 の印刷イメージに重畳して第 2 の印刷イメージを生成することを特徴とする文書画像処理方法。

【請求項 1 8】 構造化文書の論理構造とレイアウト情報の組み合わせを同定する組み合わせ情報が付加情報として重畳された印刷イメージを読み取る文書画像処理方法において、前記印刷イメージを読み取って画像情報に変換し、前記画像情報から前記付加情報を読み出して前記付加情報から前記組み合わせ情報を取得し、予め設定されている前記論理構造と前記レイアウト情報の定義に基づいて前記構造化文書の文書要素の意味もしくは構造を識別するための識別子と前記文書要素が前記印刷イメージ内に占める領域情報の対応関係を得て前記識別子と前記識別子に対応する前記領域情報を取得し、前記画像情報のレイアウトを認識して文字情報の含まれる領域である文字ブロックを抽出し、前記領域情報に対応する前記文字ブロック内の文字を認識してテキストデータを取得し、前記論理構造を得るとともに前記識別子に対応する文書要素位置に前記テキストデータを挿入して前記構造化文書を得ることを特徴とする文書画像処理方法。

【請求項 1 9】 論理構造を持った構造化文書にレイアウトを施して印刷可能とする文書画像処理をコンピュータに実行させるプログラムを格納したコンピュータ読取可能な記憶媒体において、論理構造に対応する体裁情報を定義したスタイルデータに基づいて構造化文書にレイアウトを施して印刷処理に適合した第 1 の印刷イメージに成形するとともに前記構造化文書の文書要素の意味もしくは構造を識別するための識別子と前記文書要素が前記第 1 の印刷イメージ内に占める領域情報を生成する成形処理と、前記論理構造を表す論理構造情報に加えて前記構造化文書の文書要素の意味もしくは構造を識別するための識別子と前記文書要素が前記第 1 の印刷イメージ内に占める領域情報の対応関係を示す付加情報を生成する付加情報作成処理と、前記第 1 の印刷イメージに前記付加情報を重畳して第 2 の印刷イメージを生成する付加情報重畳処理をコンピュータに実行させるプログラムを格納したコンピュータ読取可能な記憶媒体。

【請求項 2 0】 構造化文書の論理構造を表す論理構造情報に加えて前記構造化文書の文書要素の意味もしくは構造を識別するための識別子と前記文書要素が印刷イメージ内に占める領域情報の対応関係を示す付加情報が重

置された印刷イメージを読み取る文書画像処理をコンピュータに実行させるプログラムを格納したコンピュータ読取可能な記憶媒体において、前記印刷イメージを読み取り画像情報に変換する読み取り処理と、前記画像情報から前記付加情報を読み出す付加情報読み出し処理と、前記付加情報から前記論理構造情報を得て前記論理構造を復元するとともに前記構造化文書の文書要素の意味もしくは構造を識別するための識別子と前記文書要素が前記印刷イメージ内に占める領域情報の対応関係を得て前記識別子と前記識別子に対応する前記領域情報を出力する構造復元処理と、前記画像情報のレイアウトを認識して文字情報の含まれる領域である文字ブロックを抽出して前記領域情報に対応する前記文字ブロックを出力するレイアウト解析処理と、前記文字ブロック内の文字を認識してテキストデータを出力する文字認識処理と、前記論理構造を得るとともに前記識別子に対応する文書要素位置に前記テキストデータを挿入して構造化文書を得るタグ設定処理をコンピュータに実行させるプログラムを格納したコンピュータ読取可能な記憶媒体。

【請求項 21】 論理構造を持った構造化文書にレイアウトを施して印刷可能とする文書画像処理をコンピュータに実行させるプログラムを格納したコンピュータ読取可能な記憶媒体において、前記論理構造に対応してあらかじめ定義されたスタイルデータに基づいて構造化文書にレイアウトを施して第 1 の印刷イメージに成形するとともに前記論理構造と前記レイアウト情報の組み合わせを同定する組み合わせ情報を生成する成形処理と、前記組み合わせ情報を付加情報とする付加情報作成処理と、前記第 1 の印刷イメージに前記付加情報を重畳して第 2 の印刷イメージを生成する付加情報重畳処理をコンピュータに実行させるプログラムを格納したコンピュータ読取可能な記憶媒体。

【請求項 22】 構造化文書の論理構造とレイアウト情報の組み合わせを同定する組み合わせ情報が付加情報として重畳された印刷イメージを読み取る文書画像処理をコンピュータに実行させるプログラムを格納したコンピュータ読取可能な記憶媒体において、前記印刷イメージを読み取り画像情報に変換する読み取り処理と、前記画像情報から前記付加情報を読み出す付加情報読み出し処理と、前記付加情報から前記組み合わせ情報を得て予め設定されている前記論理構造と前記レイアウト情報の定義に基づいて前記構造化文書の文書要素の意味もしくは構造を識別するための識別子と前記文書要素が前記印刷イメージ内に占める領域情報の対応関係を得て前記識別子と前記識別子に対応する前記領域情報を出力する構造復元処理と、前記画像情報のレイアウトを認識して文字情報の含まれる領域である文字ブロックを抽出し前記領域情報に対応する前記文字ブロックを出力するレイアウト解析処理と、前記文字ブロック内の文字を認識してテキストデータを出力する文字認識処理と、前記論理構造

を得るとともに前記識別子に対応する文書要素位置に前記テキストデータを挿入して前記構造化文書を得るタグ設定処理をコンピュータに実行させるプログラムを格納したコンピュータ読取可能な記憶媒体。

【発明の詳細な説明】

【0001】

【発明の属する技術分野】本発明は、構造化文書をプリントする際に論理構造とレイアウト構造に関する情報を可視または不可視に埋め込み、また、そのような論理構造とレイアウト構造に関する情報が埋め込まれた文書の再入力の際に読み出された論理構造とレイアウト構造を用いて構造化文書を生成する文書画像処理装置、文書画像処理方法、および、そのような文書画像処理をコンピュータに実行させるプログラムを格納したコンピュータ読取可能な記憶媒体に関するものである。

【0002】

【従来の技術】電子的に作成された文書の異機種間での交換、再利用、検索、コンピュータによるデータの解釈などの用途に、構造化文書の利用が期待されている。代表的な構造化文書としては、ISO8879の標準一般化マーク付け言語（SGML：Standard Generalized Markup Language）が知られている。また、近年では、SGMLのサブセットにインターネット向けの機能を追加したXML（eXtensible Markup Language）が注目を集めており、企業間の電子商取引の交換フォーマットとしても期待されている。

【0003】構造化文書では、例えば、文書の表題、章、節、段落などの文書要素を識別するためにタグと呼ばれる符号が用いられる。タグは、文書の要素を識別する名称を“<”と“>”で囲んだものであり、文書要素の前後に開始タグ（<タグ名>と表記）と、終了タグ（</タグ名>と表記）が挿入される。開始タグと終了タグの間には、別の開始タグと終了タグが存在する入れ子構造となることもある。このように、構造化文書では、タグを用いることによって、文書の構造や文書を構成する要素の意味を表現することができる。

【0004】タグは、文書の内容によって自由に設定できるが、文書の作り手と受け手で相互に理解できるものでなければならない。このため、使用されるタグの集合が定義されており、この定義を文書型と呼んでいる。SGMLやXMLではDTD（Document Type Definition）と呼ばれる。XML文書では、DTDを別ファイルとしてファイル名で指定したり、XML文書中に記述することができる。

【0005】図8は、文書型の記述の概要を示す説明図である。図8に示す各行の記述は要素型宣言と呼ばれ、例えば1行目では「文書」が「表題」と「本文」から構成されることを示している。また、2行目では、「表題」がテキストであることを示している。ここで“#P

C DATA”は、XMLではテキストを表すキーワードとして用いられる。さらに、3行目では、「本文」が「第1章」、「第2章」などから構成されていることを示している。このようにして文書中に現れるすべてのタグとともに文書の論理的な構造が定義される。

【0006】一方、構造化文書は、タグ付けされた文書要素の羅列であり、レイアウトに関する情報は一切持っていない。このため、構造化文書をディスプレイ表示したり、印刷するためには、レイアウトやフォント種別、フォントサイズなどの文書の体裁を指定したスタイルデータを与える必要がある。

【0007】図9は、構造化文書を表示あるいは印刷する場合の構成例を示す説明図である。図中、31はフォーマッタである。フォーマッタ31は、入力される構造化文書に対し、与えられたスタイルファイルに定義されているスタイルデータ（レイアウト情報、フォント、文字サイズ等）を参照してレイアウトを行う。この時、構造化文書の持つ論理構造に対応付けられた文書の要素は、ページ内の2次元的な領域の配置（レイアウト構造）に対応づけられ、画面表示や印刷に用いられるページイメージ（以下、印刷イメージと呼ぶ）が生成される。

【0008】図10は、構造化文書のフォーマットの一例の説明図である。図10（A）には、タグによって表現された構造化文書の一例を示している。このような構造化文書をフォーマット処理することによって、例えば図10（B）に示すようなページレイアウトを持つ印刷イメージが得られる。

【0009】構造化文書に特定のレイアウト構造を与えてディスプレイに表示している場合は、論理構造と表示内容の関連を保持することは可能である。しかし、一旦プリント出力されたり、表示イメージが画像として取り込まれると、論理構造は失われてしまう。再び論理構造を付与するためには、文書要素を選択して対応するタグを選択するなど、労力が大きかった。

【0010】紙原稿のレイアウト構造に基づいて、論理構造を抽出する技術としては、例えば特開平10-49522号公報に開示される「文書認識装置」（以下、文献1と記す）がある。この技術は、OCRの認識結果から文書の構造上の特徴を把握し、DTD（Document Type Definition：文書型定義情報）を自動的に作成する。もしくは既にあるDTDで利用可能なものを利用する。例えば、原稿を走査して、用紙の最初に現れる中央寄せされた語句は用紙の表題、と判断するなど、あらかじめ文書のレイアウトを想定したルールを持ち、これに従ってDTDを作成するものである。

【0011】また同様の技術として、例えば特開平11-328306号公報に開示される「文書画像の論理要素抽出方法、装置および記録媒体」（以下、文献2と記

す）がある。これは、あらかじめ複数のサンプル画像を用いて、論理構造と対応付けられたレイアウトの論理モデルを作成する。入力された文書のレイアウト構造を抽出して、論理モデルとの特徴マッチングを行い、もっとも近い論理モデルを用いて入力文書の書誌的事項を抽出するものである。

【0012】これら文献1または文献2に記載されているような従来技術では、あらかじめ対象原稿のレイアウトが既知である必要があり、想定外の前稿が入力された場合には、論理構造の抽出ができないという問題があった。

【0013】一方、紙文書を入力して、レイアウトや文字の認識を行う場合に、その認識精度を高めるため、紙文書中に可視もしくは不可視の情報を埋め込んでおく技術として、例えば特開平7-168912号公報に開示される「プリント文書の光学的文字認識能力強化方法及びその手段」（以下、文献3と記す）がある。これは、レイアウト解析、OCRの結果に対してエラーの発生や発生場所を検出したり、エラーを訂正するための情報を文書マーカーとして記録しておくものである。文書マーカーを不可視に埋め込むためには、ロゴに隠す、不可視のインクを用いる、フォーマットへの隠蔽などの手法が用いられている。

【0014】また、レイアウト時の位置情報を不可視の情報として埋め込んでおく技術として、例えば特開平2000-196857号公報に開示されている「印刷装置、情報読取装置、画像処理装置、画像処理システム、情報読取方法、及び記憶媒体」（以下、文献4と記す）がある。これは、社名、住所、電話番号などの情報が印刷されている位置情報を電子透かしとして印刷イメージに埋め込んでおき、読取時に位置情報を読み出してその位置から各情報のイメージを取得し、文字認識するものである。

【0015】しかし、これら文献3あるいは文献4に記載の従来技術は、埋め込まれる情報に論理構造は含まれておらず、また、紙文書を再入力した時に再び構造化文書を得ることを目的とするものではなかった。

【0016】以上述べたように、従来技術においては、構造化文書を紙に出力する際のレイアウトを自由にする、紙文書を再入力した時に再び論理構造を持つ構造化文書を得ることが困難であった。また、文書を再入力する際の認識精度を高めるために、文書に付加情報を可視、もしくは不可視に埋め込む技術はあったが、文書の論理構造に着目したものはなかった。

【0017】

【発明が解決しようとする課題】本発明は、上述した事情に鑑みてなされたもので、構造化文書を印刷可能なイメージとする際に、論理構造とレイアウト構造に関する情報を、可視もしくは不可視の情報として埋め込んでおく。そして、文書のイメージを再入力する際に、これら

の埋め込まれた情報を用いて構造化文書を得ることのできる、文書画像処理装置および文書画像処理方法、並びに記憶媒体を提供することを目的とするものである。

【0018】

【課題を解決するための手段】本発明は、論理構造を持った構造化文書にレイアウトを施して印刷可能とする際には、論理構造に対応する体裁情報を定義したスタイルデータに基づいて構造化文書にレイアウトを施して印刷処理に適合した第1の印刷イメージに成形するとともに、構造化文書の文書要素の意味もしくは構造を識別するための識別子と該文書要素が第1の印刷イメージ内に占める領域情報を生成する。そして、論理構造を表す論理構造情報に加えて、構造化文書の文書要素の意味もしくは構造を識別するための識別子と前記文書要素が第1の印刷イメージ内に占める領域情報の対応関係を示す付加情報を生成して、付加情報を第1の印刷イメージに重畳し、第2の印刷イメージを生成することを特徴とするものである。

【0019】また、このような構造化文書の論理構造を表す論理構造情報に加えて前記構造化文書の文書要素の意味もしくは構造を識別するための識別子と該文書要素が印刷イメージ内に占める領域情報の対応関係を示す付加情報が重畳された印刷イメージを読み取る際には、印刷イメージを読み取って画像情報に変換し、画像情報から付加情報を読み出し、付加情報から論理構造情報を得て論理構造を復元する。それとともに、構造化文書の文書要素の意味もしくは構造を識別するための識別子と該文書要素が印刷イメージ内に占める領域情報の対応関係を得て、識別子と識別子に対応する領域情報を取得し、画像情報のレイアウトを認識して文字情報の含まれる領域である文字ブロックを抽出する。そして、領域情報に対応する前記文字ブロック内の文字を認識してテキストデータとし、論理構造を得るとともに識別子に対応する文書要素位置に認識したテキストデータを挿入して構造化文書を得ることを特徴とするものである。

【0020】より付加情報を減少させるための構成として、論理構造を持った構造化文書にレイアウトを施して印刷可能とする際に、論理構造に対応してあらかじめ定義されたスタイルデータに基づいて構造化文書にレイアウトを施して第1の印刷イメージに成形するとともに、論理構造と前記レイアウト情報の組み合わせを同定する組み合わせ情報を生成して付加情報とし、該付加情報を第1の印刷イメージに重畳して第2の印刷イメージを生成することができる。

【0021】また、このような構造化文書の論理構造とレイアウト情報の組み合わせを同定する組み合わせ情報が付加情報として重畳された印刷イメージを読み取る際には、印刷イメージを読み取って画像情報に変換し、画像情報から付加情報を読み出して付加情報から組み合わせ情報を取得し、予め設定されている論理構造とレイ

アウト情報の定義に基づいて構造化文書の文書要素の意味もしくは構造を識別するための識別子と文書要素が前記印刷イメージ内に占める領域情報の対応関係を得て、識別子と識別子に対応する前記領域情報を取得する。そして、画像情報のレイアウトを認識して文字情報の含まれる領域である文字ブロックを抽出し、領域情報に対応する文字ブロック内の文字を認識してテキストデータを取得し、論理構造を得るとともに識別子に対応する文書要素位置に認識したテキストデータを挿入して、構造化文書を得るように構成することができる。

【0022】このような構成によって、構造化文書にレイアウトを施した印刷イメージに付加情報を重畳して例えば印刷し、その印刷物のイメージを読み取って付加情報をもとに論理構造を復元して構造化文書を取得することができる。

【0023】

【発明の実施の形態】図1は、本発明の第1の実施の形態を示す構成図である。図中、1はフォーマッタ、2はスタイルデータ記憶部、3はスタイル指示部、4は付加情報作成部、5は付加情報重畳部、6は印字部、7は原稿読み取り部、8は補正処理部、9は付加情報読み出し部、10は構造復元部、11はレイアウト解析部、12はOCR部、13はタグ設定部である。図1には、構造化文書の論理構造を含む付加情報が重畳された印刷イメージを印刷する側の構成と、そのような印刷物を読み取って構造化文書を復元する側の構成の両方を示している。

【0024】構造化文書を印刷する側の構成としては、フォーマッタ1、スタイルデータ記憶部2、スタイル指示部3、付加情報作成部4、付加情報重畳部5、印字部6等を有している。フォーマッタ1は、入力される構造化文書に対し、スタイルデータ記憶部2から与えられたスタイルデータ（レイアウト情報、フォント、文字サイズ等）に従ってレイアウトを行う。この時、構造化文書の持つ論理構造に対応付けられた文書の要素は、ページ内の2次元的な領域の配置（レイアウト構造）に対応づけられ、第1の印刷イメージが生成される。スタイルデータ記憶部2は、1ないし複数のスタイルデータを記憶しており、スタイル指示部3から指示されたレイアウトデータをフォーマッタ1に出力する。スタイル指示部3は、利用者がスタイルデータ記憶部2に記憶されているスタイルデータの選択及び編集を行う。付加情報作成部4は、フォーマッタ1から出力される第1の印刷イメージ内の2次元の領域情報と、構造化文書の文書型に定義されるタグの対応を所定の形式に記述して付加情報を作成する。付加情報重畳部5は、フォーマッタ1から出力される第1の印刷イメージに、付加情報作成部4で作成した付加情報を重畳して第2の印刷イメージを生成する。印字部6は、付加情報重畳部5で付加情報を重畳した第2の印刷イメージを記録紙に印字する。

【0025】また構造化文書を復元する側の構成としては、原稿読み取り部7、補正処理部8、付加情報読み出し部9、構造復元部10、レイアウト解析部11、OCR部12、タグ設定部13等を有している。原稿読み取り部7は、印字部6によって印刷された紙文書の画像を読み取る。補正処理部8は、原稿読み取り部7で読み取られた画像情報に対して、スキュー補正や、天地補正などの補正処理を行う。付加情報読み出し部9は、画像情報に重畳されている付加情報を読み出す。構造復元部10は、付加情報読み出し部9によって読み出された付加情報から、論理構造とレイアウト構造に関する情報を復元する。レイアウト解析部11は、読み取られた原稿のレイアウトを解析して、構造復元部10が出力するレイアウト構造を参照して文字ブロックを切り出す。OCR部12は、切り出された文字ブロックごとにOCR（文字認識）を施し、テキスト情報を出力する。タグ設定部13は、OCR部12から出力されたテキスト情報を、構造復元部10から出力されるタグの対応する箇所に挿入し、構造化文書を出力する。

【0026】図2は、タグ設定部13の一例を示す構成図である。図中、21は論理構造保持部、22はタグ記憶部、23はタグ開始終了判定部、24は切替部である。論理構造保持部21は、構造復元部10で復元された文書の論理構造（タグの入れ子構造）を記憶する。タグ記憶部22は、構造復元部10からレイアウト解析部11に領域情報を送ると同期して出力されるタグ情報を保持する。タグ開始終了判定部23は、論理構造保持部21から順次出力される文書の論理構造からタグ記憶部22に記憶されるタグの開始と終了を検出して切替部24の入力を切り替える。切替部24は、タグ開始終了判定部23の指示により論理構造保持部21から出力されるタグとOCR部12から出力されるテキストデータの何れかを切り替えて出力し、文書構造中にテキストデータを挿入する。

【0027】本発明の第1の実施の形態における動作について説明する。構造化文書はフォーマッタ1に入力される。なお、図1においては構造化文書と文書型が別ファイルとして構成されているものとして示しており、文書型は付加情報作成部4によって読み込まれるとともに、構造化文書内においてファイル名などで指定され、参照される。一方、スタイル指示部3において利用者が構造化文書を印刷する際に用いるスタイルデータの選択指示を行い、指示されたスタイルデータがスタイルデータ記憶部2からフォーマッタ1に入力される。なお、スタイルデータは、スタイルデータ指示部3によって利用者が設定を変更することができる。

【0028】フォーマッタ1は、スタイルデータ記憶部2に記憶されている選択されたスタイルデータに基づいて、入力された構造化文書から印刷イメージを形成する。そして、構造化文書の文書型と、文書要素に対応づ

けられたタグとこの文書要素が印刷イメージ内に占める2次元領域の対応を示す情報を付加情報作成部4に送るとともに、フォーマットされた印刷イメージを付加情報重畳部5に送る。

【0029】図3は、印刷イメージ内のレイアウトを構成する2次元領域の設定方法と2次元領域の記述方法の一例の説明図である。例えば文書作成ソフトウェアのテキストボックスなどのように、文書画像のレイアウトでは矩形領域を基本とすることが多い。また、文書認識の前処理としてレイアウト解析を行う場合には、原稿内の文字の領域を文字ブロックと呼ばれる矩形の領域として抽出することが多い。このため、本発明では文書画像のレイアウトは、矩形領域の2次元の配置情報と考える。

【0030】図3（A）は、図10（B）に示したページレイアウトを持つ印刷イメージに対し、印字要素に外接する矩形領域を設定したものである。それぞれの矩形領域の各辺が用紙端と平行であるものとし、原稿の最左上端を原点とし、水平方向をx軸、垂直方向をy軸とする座標系を設定する。これにより、印刷イメージ内の矩形領域は、最左上の角を開始座標（x, y）、矩形領域の幅（x軸方向）をw、矩形領域の高さ（y軸方向）をhで表現することができる。図3（A）に示すページレイアウト内に表示される文書要素に対応する2次元領域は、図3（B）に示すように記述される。レイアウトを構成する矩形領域の位置と大きさは、スタイルデータに基づいてフォーマッタ1によって決定される。矩形領域の位置と大きさが、原稿内の座標系に対して固定である場合と、例えば、フォントサイズや段落幅の変更に対応して変更される場合がある。

【0031】なお、本発明においては、矩形領域の記述方法はこれに限定されるものではなく、矩形領域の4端点の座標を用いるなど他の記述方法を用いても構わない。また、座標系の単位についても限定されるものではなく、ミリメートル単位やインチ単位、その他の単位を用いてもかまわない。また、原稿サイズで正規化した座標系の単位を用いてもかまわない。

【0032】付加情報作成部4では、文書型と、タグと印刷イメージ内の2次元領域の対応関係を所定の記述形式で生成し、付加情報として付加情報重畳部5に送る。

図4は、タグと、印刷イメージ内の2次元領域の対応の記述の一例の説明図である。最も単純な記述方法は、タグの名称と矩形領域の情報を対で記述する方法である。同じ文書内に同一の名称を持つタグが存在する場合は、上位の階層のタグから記述することでそれぞれを識別できるようにする。例えば、〈段落1〉のタグだけではどの章の段落かを判別できない場合には、上位の階層を示す〈第1章〉のタグとともに記述して区別すればよい。図4においては、階層を「/」で区切って示している。

【0033】図5は、タグと、印刷イメージ内の2次元領域の対応の記述の別の例の説明図である。タグと、印

刷イメージ内の2次元領域の対応の記述として、例えば図5に示すように記述することもできる。この例では、図10(A)に示したようにタグを用いて表現された文書の論理構造において、開始タグと終了タグに囲まれる部分から文書要素を取り除き、代わりに対応する矩形領域の情報を埋め込むものである。このような操作によって、例えば図5(A)に示すような記述を得ることができる。この場合には、先の例のように同一名称のタグが存在する場合でも、一対一にレイアウト構造との対応づけを表現することができる。

【0034】ただし、図5(A)に示す記述では、文書型とともに付加情報として追加するには、表現が冗長である。このため、図5(A)に示した表記に対し、例えば、開始タグと終了タグをそれぞれ開き括弧と閉じ括弧で置き換え、タグ名称を除去することによって、図5

(B)に示すような記述を得ることができる。さらに、改行やインデントは、人が見る時の理解を容易にするためのものであり、これらはタグ(図5(B)では括弧に置き換えられている)の入れ子関係から階層構造を把握する上では不要であり、実際に埋め込まれる情報からは除外することができる。このようにして、図5

(C)に示す記述が得られる。この記述を文書型とともに付加情報とすることができる。なお、この例では階層構造の表記に丸括弧を用いたが、これに限られるものではなく、種々の表記が可能である。

【0035】付加情報重畳部5では、付加情報作成部4で作成された、論理構造とレイアウト構造の対応の記述を、フォーマッタ1によって生成された印刷イメージに付加情報として重畳する。印刷イメージへの情報の重畳方法としては、人の目に見える形式(可視)で埋め込む場合と、人の目には見えない、もしくは、見えづらい形式(不可視)で埋め込む場合がある。文書に可視に情報を埋め込む技術としては、バーコードやグリフコードを用いるなどの公知の手法を適用することができる。なお、可視に情報を埋め込む場合には、あらかじめフォーマッタ1において印刷イメージを形成する際に、用紙内に情報を埋め込むための領域を確保しておくことよい。

【0036】一方、文書に不可視に情報を埋め込む技術としては、例えば特開2000-138813号公報に開示される「画像処理装置および画像処理方法」(以下、文献5と記す)などの技術が適用可能である。この他にも、不可視なインク、トナー等を用いて情報を埋め込む方法等も適用可能である。

【0037】付加情報重畳部5によって論理構造とレイアウト構造の対応の記述が埋め込まれた印刷イメージは、印字部6によって用紙に記録される。以上の構成と動作により、構造化文書にその構造情報を含む付加情報を重畳して印刷することができる。

【0038】次に、このようにして印刷された紙原稿を

してゆく。原稿読み取り部7は、紙原稿を走査してラスト画像データを生成し、補正処理部8においてラスト画像データに対する補正処理を実行し、補正されたラスト画像データを付加情報読み出し部9に入力する。補正処理の内容としては、紙原稿が傾いて読み込まれた場合のスキュー補正や、天地が逆転して読み込まれた場合の天地補正などがあり、それぞれについての公知の手法を適用することができる。

【0039】付加情報読み出し部9は、紙原稿に重畳されている付加情報を読み出す。読み出す方法は、付加情報重畳部5における情報の重畳方法に対応した方法を用いればよい。例えばバーコードやグリフコードによって情報が可視に埋め込まれている場合は、これらの情報を読み取って元の情報に復元する公知の手法を利用すればよい。また、上述の文献5に示されている手法によって不可視に情報が埋め込まれている場合についても、同文献に記載されている読み出し方法によって読み出せばよい。この他、不可視なインク、トナー等を用いて情報が埋め込まれた場合についても同様である。付加情報読み出し部9によって読み出された付加情報は構造復元部10に入力され、それ以外のラスト画像情報はレイアウト解析部11に入力される。

【0040】構造復元部10では、論理構造と、論理構造とレイアウト構造の構成要素の対応関係が復元される。上述のように、付加情報としては、文書型と、タグと印刷イメージ内の2次元領域の対応関係が得られる。タグと印刷イメージ内の2次元領域の対応関係としては、例えば図4に示すようなデータ、あるいは図5

(B)または(C)に示すようなデータが得られるので、これらの論理構造の情報と文書型とを対比することによって論理構造を復元することができる。図6は、復元された論理構造の一例の説明図である。図6は、図4あるいは図5(B)又は(C)に示した論理構造とレイアウト構造の対応関係に基づいて、文書型から論理構造を復元した例を示している。図4に示した論理構造とレイアウト構造の対応関係の情報を用いる場合には、それぞれのタグ名が含まれているので、文書型における構成要素との対応付けを行うことによって論理構造を復元することができる。また、図5(B)又は(C)に示した論理構造とレイアウト構造の対応関係の情報を用いる場合には、論理構造の入れ子関係が分かるので、文書型との対応付けによって各タグ名を特定することができる。なお、構造復元部10において文書型に基づいて復元された論理構造は、図6に示しているようにタグの階層構造だけを持つもので、内容は含まれない。この論理構造は、タグ設定部13に送られる。

【0041】また、構造復元部10は、レイアウト構造を構成する矩形領域の記述情報を、順次、レイアウト解析部11に送り、OCR部12により認識処理を行う文字ブロックを通知するとともに、対応するタグをタグ設

定部 13 に通知する。

【0042】一方、レイアウト解析部 11 では、ラスタ画像データ中から、文字ブロックや写真などの絵柄領域を、矩形領域を単位として検出する。これらの矩形領域の内、構造復元部 10 から受け取った矩形領域情報に対応するラスタ画像データの部分領域を OCR 部 12 に送り、OCR 部 12 によって文字認識処理が施される。得られたテキストコード情報がタグ設定部 13 に送られる。

【0043】タグ設定部 13 では、文書型などによりあらかじめ設定されている論理構造の雛形に対し、構造復元部 10 より通知されるタグに対応する位置に、OCR 部 12 から送られるテキストデータを設定して構造化文書を復元する。

【0044】このタグ設定部 13 における動作を、図 2 を用いて説明する。はじめに、構造復元部 10 で復元された文書の論理構造は、タグ設定部 13 内の論理構造保持部 21 に記憶される。次に構造復元部 10 によって、レイアウト構造に対応を持つタグの内、論理構造内で最も先頭に近いタグがタグ記憶部 22 に設定される。これに同期して、構造復元部 10 では設定されたタグに対応する領域情報をレイアウト解析部 11 に送出する。

【0045】タグ開始終了判定部 23 は、論理構造保持部 21 から順次出力されるタグを監視し、タグ記憶部 22 に記憶されているタグの開始タグが検出されるまで、論理構造保持部 21 の出力がそのまま出力されるように切替部 24 を制御する。タグ開始終了判定部 23 が論理構造保持部 21 の出力からタグ記憶部 22 に記憶されるタグの開始タグを検出すると、この開始タグの出力後に切替部 24 の入力を OCR 部 12 の出力に切り替える。OCR 部 12 では、構造復元部 10 からレイアウト解析部 11 に出力された領域情報に対応する読み取り画像内の文字認識処理が行われてテキストデータとして出力される。このテキストデータが切替部 24 に入力され、そのまま出力されることになる。

【0046】OCR 部 12 によるテキストデータの出力終了が図示しない制御信号によってタグ開始終了判定部 23 に通知され、切替部 24 の入力を論理構造保持部 21 に切り替える。論理構造保持部 21 からは、タグ記憶部 22 に記憶されているタグの終了タグが出力される。この時、構造復元部 10 からは、レイアウト構造に対応を持つタグの内、前記論理構造内で先頭から 2 番目のタグがタグ記憶部 22 に設定される。これに同期して、構造復元部 10 では新たに設定されたタグに対応する領域情報をレイアウト解析部 11 に送出する。

【0047】以上の動作を繰り返すことにより、文書型より生成された論理構造に対し、タグに対応づけられた文書領域の OCR 結果が順次挿入されることになる。このようにして、読み取られた紙文書から構造化文書を復元することができる。

【0048】上述の例では、文書型は必ず付加情報として重畳しているが、これを簡略化することで情報量を減少することもできる。例えば、XML を企業間の電子商取引に応用する場合などでは、あらかじめ関係する企業間において取引内容に基づく文書型 (DTD) が標準化されていくと考えられる。このように、構造化文書を印刷する側と読み取る側で、文書型についての情報が共有されている場合は、必ずしも文書型を埋め込む必要はなく、文書型を特定できる情報を埋め込んだり、あるいは文書型の情報を埋め込まなくてもよい。

【0049】また、複数の構造化文書を 1 枚の用紙にまとめて出力する場合には、文書型とともに、タグと矩形領域の対応を示す情報の埋め込みを、構造化文書の数だけ繰り返せば良い。

【0050】次に、本発明の第 2 の実施の形態について説明する。情報を埋め込む場合には、埋め込まれる情報の量を考慮する必要がある。例えば、バーコードを用いて情報を可視に埋め込む場合に情報量が多いと、原稿上のバーコードを印刷する面積を大きくする必要があり、本来の文書内容の印刷領域に制限が生じることが考えられる。この本発明の第 2 の実施の形態は、このような問題に鑑みてなされたものであり、構造化文書の論理構造と印刷時のレイアウト構造を対応付ける情報をより少ない情報量で表現しようとするものである。なお、構成は上述の第 1 の実施の形態とほぼ同様であるので、上述の図 1 及び図 2 を参照するものとして相違点について主に説明してゆく。また、動作についても、同様の動作については説明を省略する。

【0051】上述の第 1 の実施の形態においては、レイアウト構造を構成する領域情報を領域の開始座標 (x, y)、幅 (w)、高さ (h) で表現していた。しかし、印刷時と読み取り時に原稿内の矩形領域に同一の識別子を割り当てることができれば、領域情報を表す情報量を削減することができる。例えば、上述の図 3

(A) に示したページレイアウトにおいて、印刷イメージを原点から走査してゆき、矩形領域の開始座標が現れる順序に 1 から 5 の識別子を割り当てることが考えられる。印刷時は、付加情報作成の際に、レイアウト構造を構成する 2 次元領域に一意に識別子として番号を割り振ることができる。また、原稿を読み取る側では、レイアウト解析により検出された矩形領域に対して、同様の手順で識別子の番号を割り振ればよい。このとき、印刷時と読み取り時で同じ矩形領域については識別子が同じになるように、識別子の割り振りを行い、またレイアウト解析を行えばよい。

【0052】図 7 は、本発明の第 2 の実施の形態におけるタグと印刷イメージ内の 2 次元領域との対応の記述の一例の説明図である。ここでは、図 7 (A)、(B) に示した 2 つの異なるレイアウト構造を考える。上述のように矩形領域の開始座標が現れる順序 (この例では水平

方向優先)に従って識別子を割り当てると、図7

(A), (B)において数字で示すように各矩形領域に識別子を割り当てることができる。

【0053】いま、図7(C)に示すような論理構造をもつ構造化文書があり、タグC1, C2, C3で定義される内容を、図7(A), (B)に示すレイアウト構造に従ってフォーマットする。これによってタグC1, C2, C3で定義される内容は、それぞれ、例えば図7(E), (F)に示すように割り当てることができる。

【0054】この時、論理構造とレイアウト構造の対応の記述は、上述の第1の実施の形態における図4に示した例と同様に、タグと識別番号を対で記述する方法がもっとも単純である。図7(G), (H)は、それぞれ図7(E), (F)に対応する。例えば図7(E)に示すようにタグC1の内容は矩形領域1に配置され、タグC2の内容は矩形領域2に配置され、タグC3の内容は矩形領域3に配置されている。この関係を図7(G)に示すように表記すれば、論理構造と矩形領域との対応を記述することができる。同様に、図7(F)に示すようにタグC1の内容は矩形領域1に配置され、タグC2の内容は矩形領域3に配置され、タグC3の内容は矩形領域2に配置されている場合には、この関係を図7(H)に示すように表記すれば、論理構造と矩形領域との対応を記述することができる。

【0055】なお、図7(G), (H)における「ID=N」の部分、図7(C)に示した文書構造がNというID(識別子)で管理されている場合の簡略表記である。この場合、印刷時と印刷物の読み取り時で同じIDで文書構造が管理されていることを前提となる。もちろん、上述の第1の実施の形態と同様に、この部分に文書型を設定してもよい。

【0056】また、やはり上述の第1の実施の形態において図5で説明した例と同様に、図7(C)に示した論理構造は、図7(D)に示すように括弧を用いた簡略表現が可能であり、タグを除去した後にタグに対応する文書内領域の識別番号を挿入して論理構造とレイアウト構造の対応を記述することができる。この場合、図7(E), (F)に対応する記述は、それぞれ図7(I), (J)に示すようになる。このような記述によって、付加情報を、文書型と、タグと識別番号の対応付けによって、より少ない情報量で簡略に記述することができる。

【0057】上述の第1及び第2の実施の形態では、構造化文書の論理構造と印刷時のレイアウト構造との組み合わせが任意である場合に適用可能である。しかし、論理構造(文書型)に対応するレイアウト構造の組み合わせが一意に定まるとともに、構造化文書を印刷する側と読み取る側においてこの組み合わせについての情報が共有される場合には、これらの組み合わせを指定する情報を付加情報として埋め込むことができる。この場合、論

理構造とレイアウト構造の対応を示す情報量を格段に少なくすることができる。

【0058】なお、上述の第1及び第2の実施の形態では、印字部6によって構造化文書及び付加情報を紙に記録する場合について述べたが、これに限らず、例えば付加情報を含む構造化文書のイメージをファイルなどに保存しておき、このイメージを読み出して構造化文書を復元する場合にも、同様にして本発明を適用可能である。

【0059】上述の各実施の形態やその変形及び組み合わせは、コンピュータプログラムによっても実現することが可能である。その場合、そのプログラムおよびそのプログラムが用いるデータなどは、コンピュータが読み取り可能な記憶媒体に記憶することも可能である。記憶媒体とは、コンピュータのハードウェア資源に備えられている読取装置に対して、プログラムの記述内容に応じて、磁気、光、電気等のエネルギーの変化状態を引き起こして、それに対応する信号の形式で、読取装置にプログラムの記述内容を伝達できるものである。例えば、磁気ディスク、光ディスク、CD-ROM、メモリカード、コンピュータに内蔵可能なメモリ等である。

【0060】

【発明の効果】以上の説明から明らかなように、本発明によれば、構造化文書を出力する際に、文書のレイアウト構造と論理構造に関する情報を、可視、または、不可視の情報として埋め込むことにより、再入力の際には埋め込まれているレイアウト構造と論理構造に関する情報に基づいて、再び構造化文書を得ることができる。また、文書のレイアウト構造と論理構造に関する情報から構造化文書を復元するので、構造化文書の出力時のレイアウト設定の自由度を大幅に向上させることができるという効果がある。

【図面の簡単な説明】

【図1】 本発明の第1の実施の形態を示す構成図である。

【図2】 タグ設定部13の一例を示す構成図である。

【図3】 印刷イメージ内のレイアウトを構成する2次元領域の設定方法と2次元領域の記述方法の一例の説明図である。

【図4】 タグと、印刷イメージ内の2次元領域の対応の記述の一例の説明図である。

【図5】 タグと、印刷イメージ内の2次元領域の対応の記述の別の例の説明図である。

【図6】 復元された論理構造の一例の説明図である。

【図7】 本発明の第2の実施の形態におけるタグと印刷イメージ内の2次元領域との対応の記述の一例の説明図である。

【図8】 文書型の記述の概要を示す説明図である。

【図9】 構造化文書を表示あるいは印刷する場合の構成例を示す説明図である。

【図10】 構造化文書のフォーマットの一例の説明図

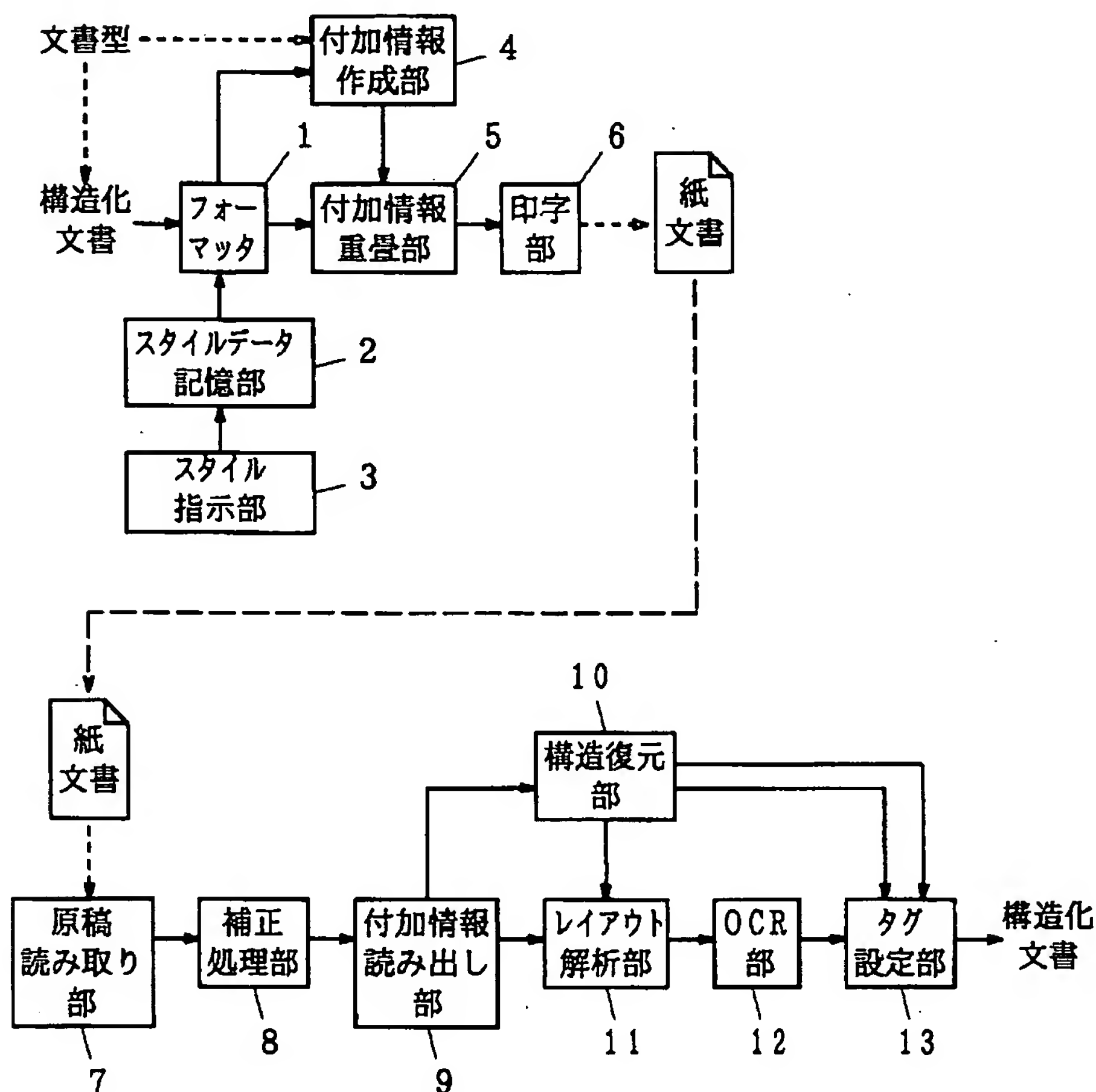
である。

【符号の説明】

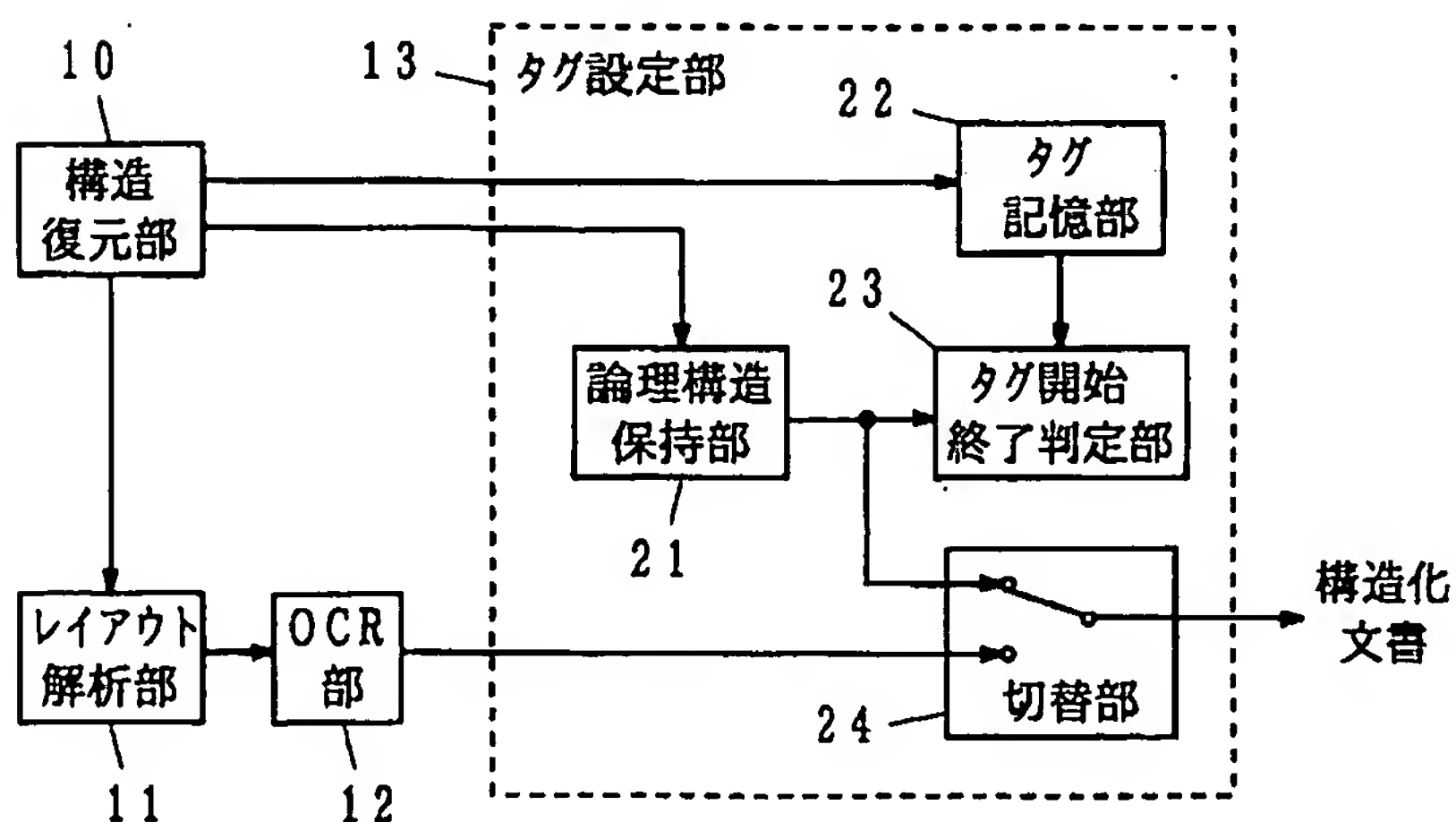
1…フォーマッタ、2…スタイルデータ記憶部、3…スタイル指示部、4…付加情報作成部、5…付加情報重畳部、6…印字部、7…原稿読み取り部、8…補正処理部、9…付加情報読み出し部、10…構造復元部、11…レイアウト解析部、12…OCR部、13…タグ設定部、21…論理構造保持部、22…タグ記憶部、23…タグ開始終了判定部、24…切替部、31…フォーマッタ。

部、9…付加情報読み出し部、10…構造復元部、11…レイアウト解析部、12…OCR部、13…タグ設定部、21…論理構造保持部、22…タグ記憶部、23…タグ開始終了判定部、24…切替部、31…フォーマッタ。

【図1】

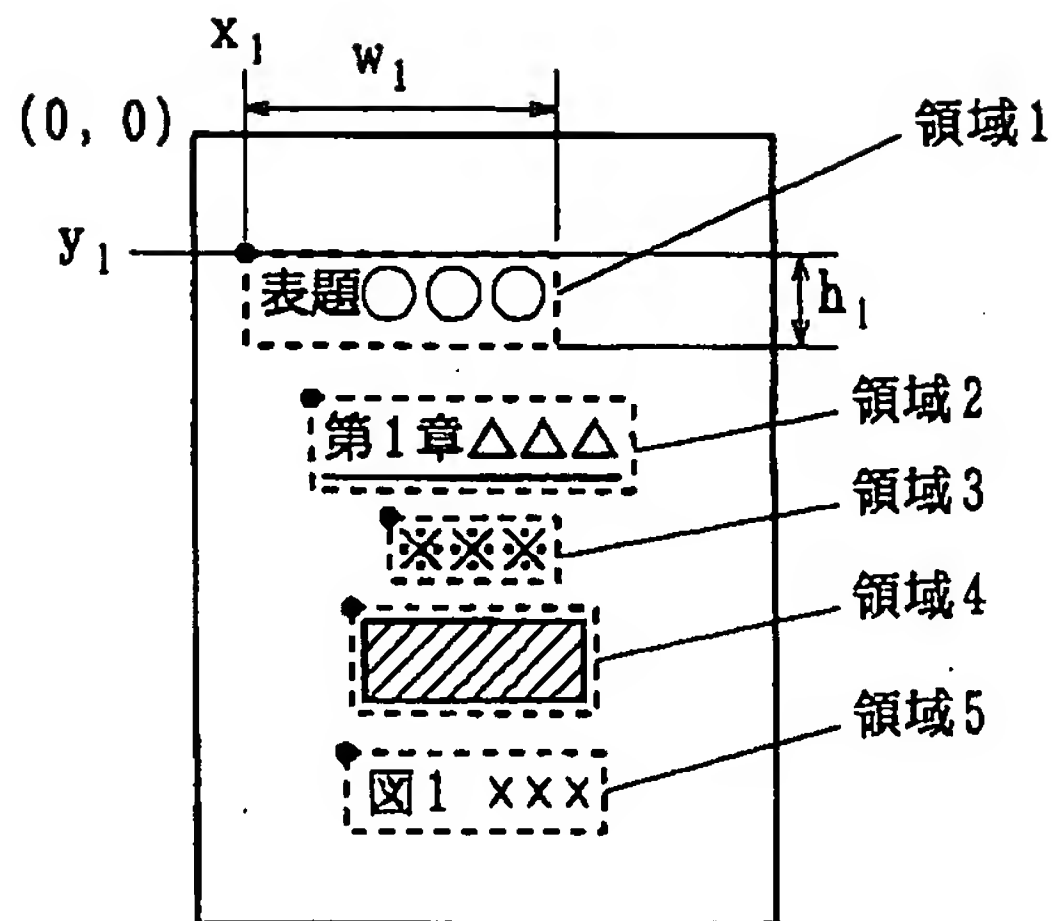


【図2】



【図3】

(A)



(B)

領域1
開始座標 (x_1, y_1) , 幅 (w_1) , 高さ (h_1)
領域2
開始座標 (x_2, y_2) , 幅 (w_2) , 高さ (h_2)
領域3
開始座標 (x_3, y_3) , 幅 (w_3) , 高さ (h_3)
領域4
開始座標 (x_4, y_4) , 幅 (w_4) , 高さ (h_4)
領域5
開始座標 (x_5, y_5) , 幅 (w_5) , 高さ (h_5)

【図4】

表題: (x_1, y_1, w_1, h_1)
章題: (x_2, y_2, w_2, h_2)
第1章/段落1: (x_3, y_3, w_3, h_3)
第1章/図/データ1: (x_4, y_4, w_4, h_4)
第1章/図/図題1: (x_5, y_5, w_5, h_5)

【図6】

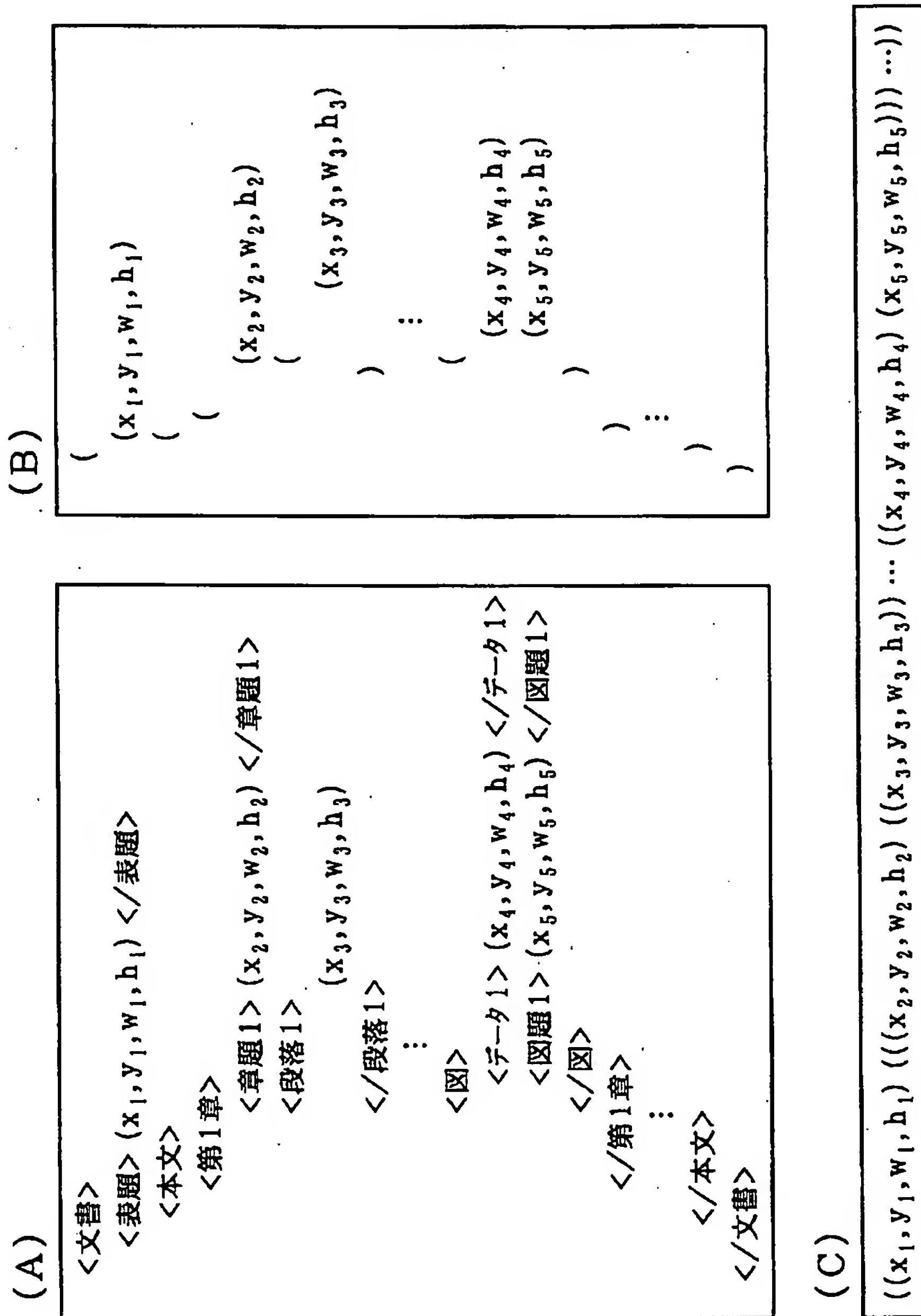
```
<文書>
  <表題></表題>
  <本文>
    <第1章>
      <章題1></章題1>
      <段落1></段落1>
      :
      <図>
        <データ1></データ1>
        <図題1></図題1>
      </図>
    </第1章>
    :
  </本文>
</文書>
```

【図8】

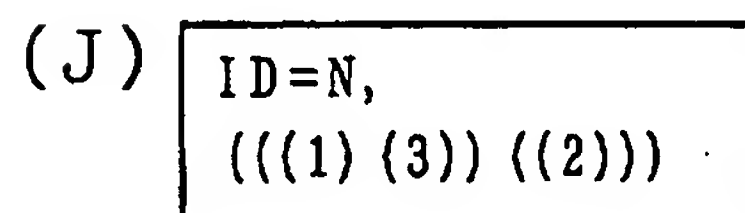
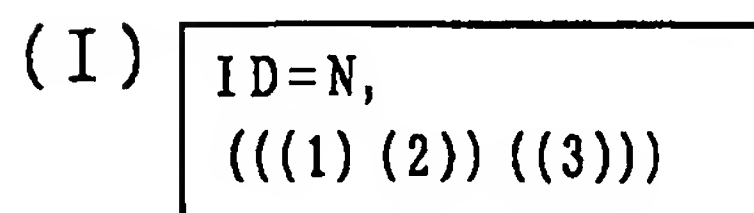
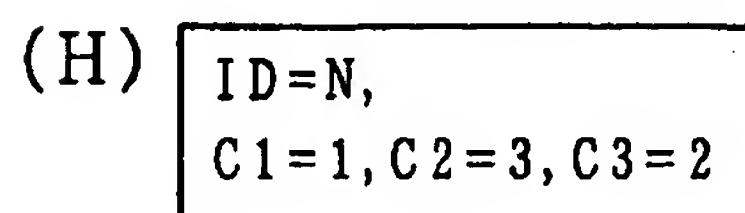
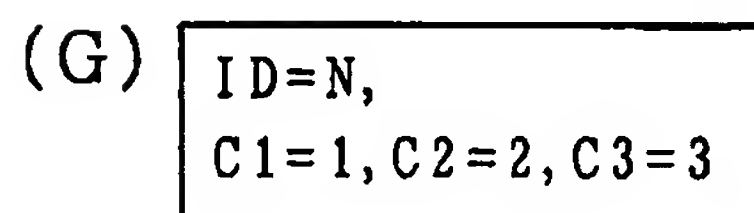
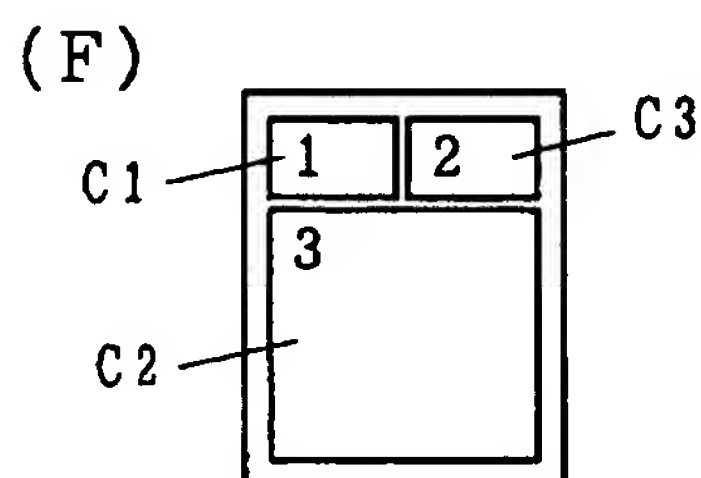
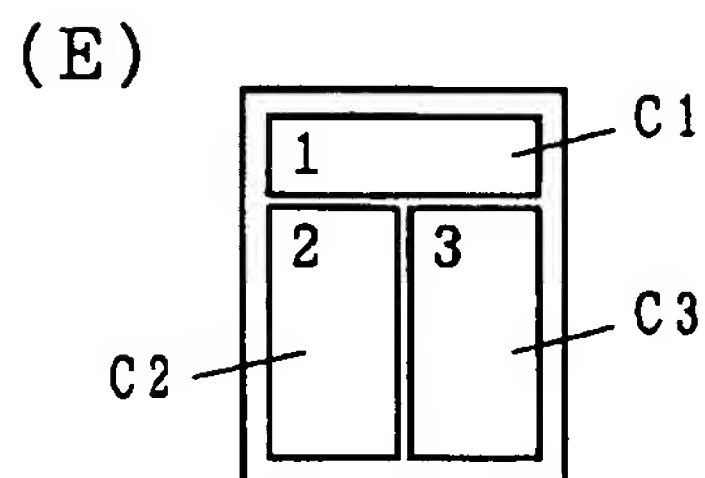
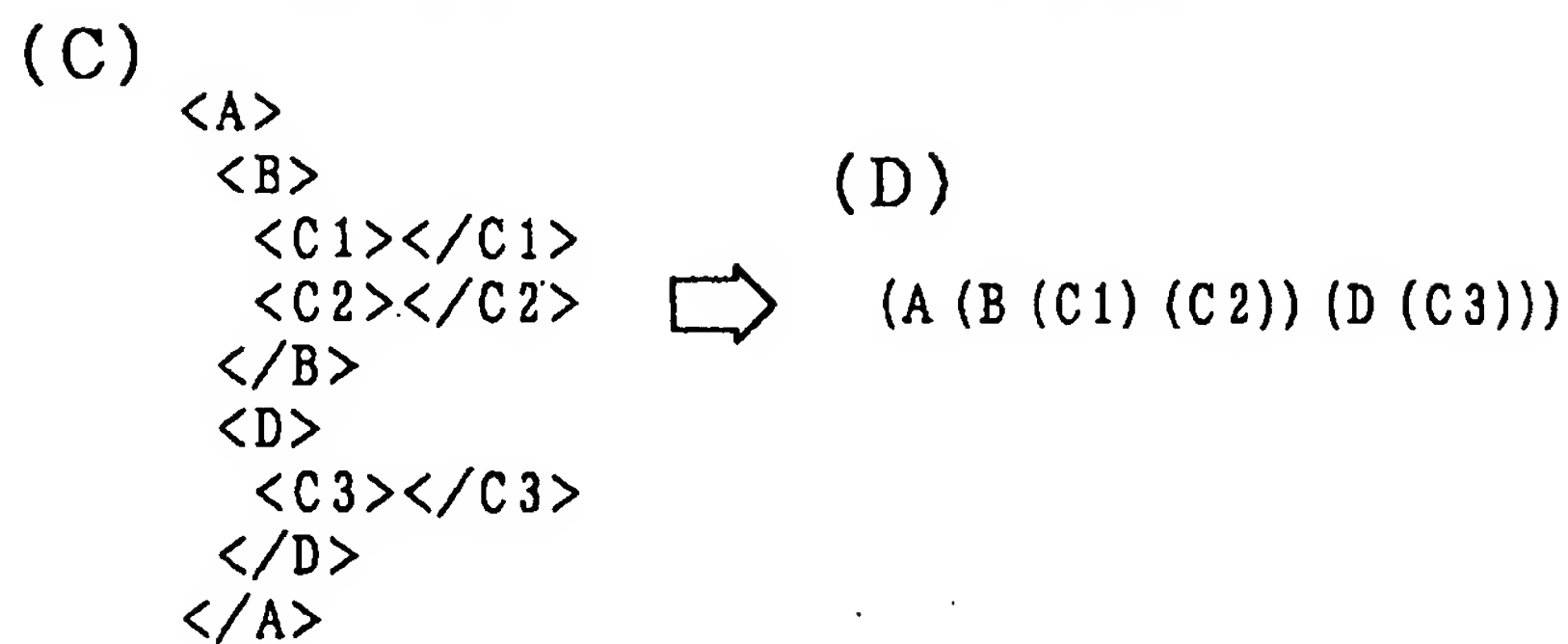
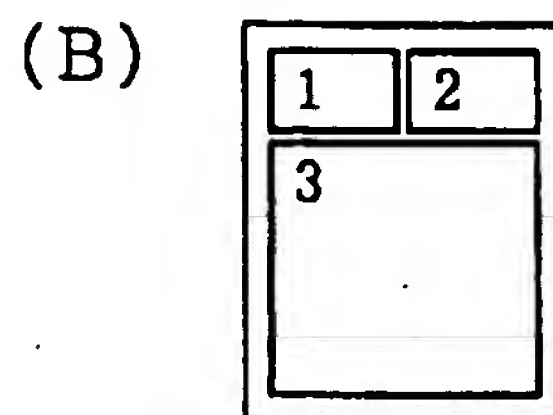
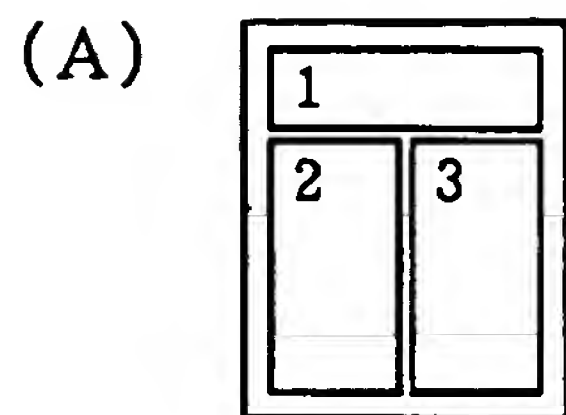
```
<!ELEMENT 文書 (表題、本文)>
<!ELEMENT 表題 (#PCDATA)>
<!ELEMENT 本文 (第1章、第2章、...)>
<!ELEMENT 第1章 (章題、段落1、...、図)>
<!ELEMENT 図 (データ、図題)>
```

⋮

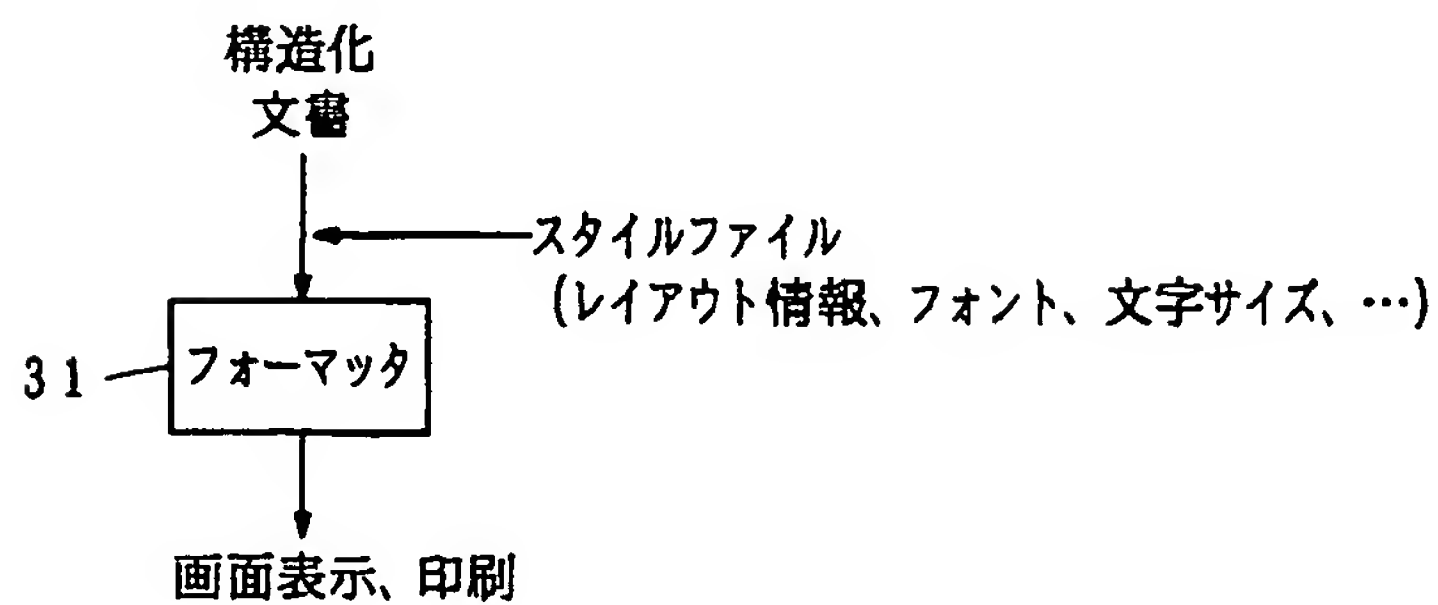
【図 5】



【図 7】

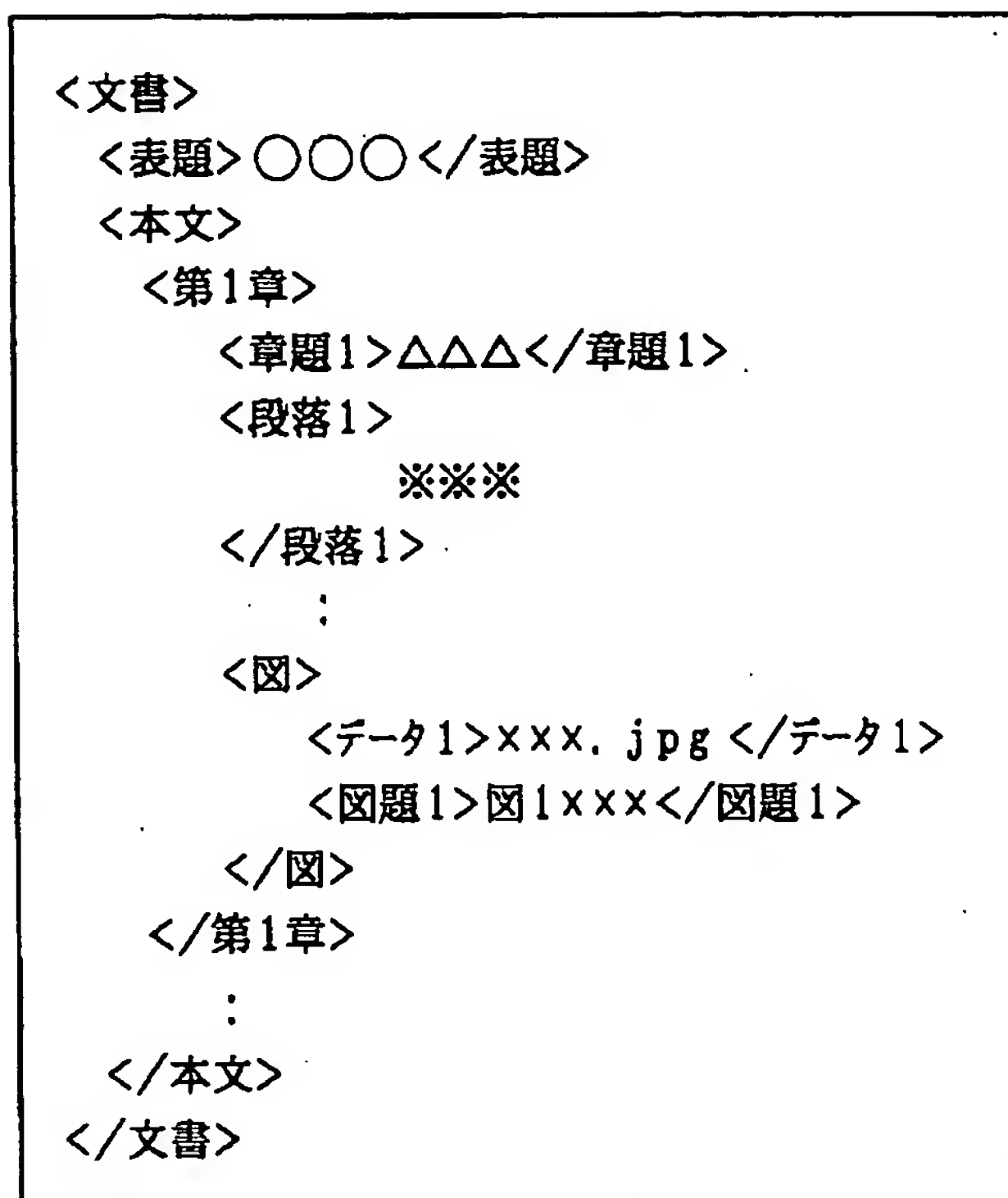


【図 9】

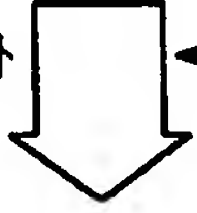


【図10】

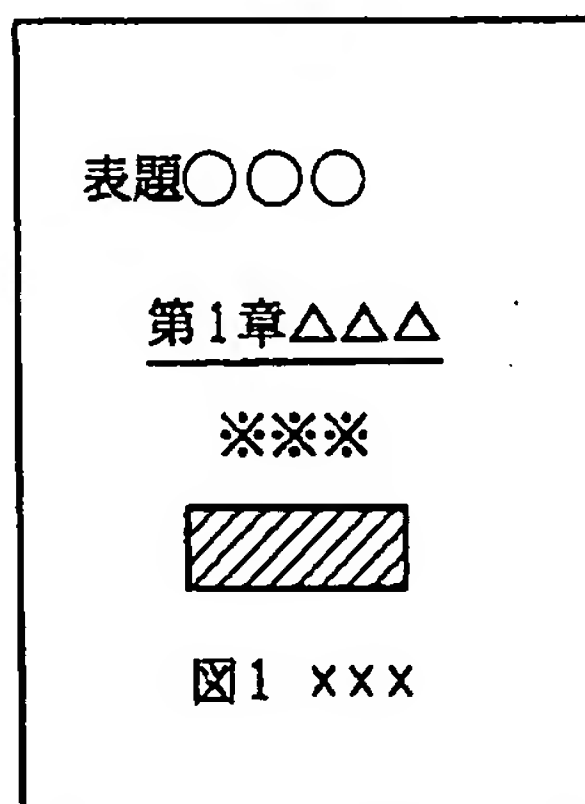
(A)



フォーマット スタイルデータ



(B)



フロントページの続き

(51) Int. Cl.⁷

H04N 1/387

識別記号

FI

H04N 1/387

フォーマット (参考)